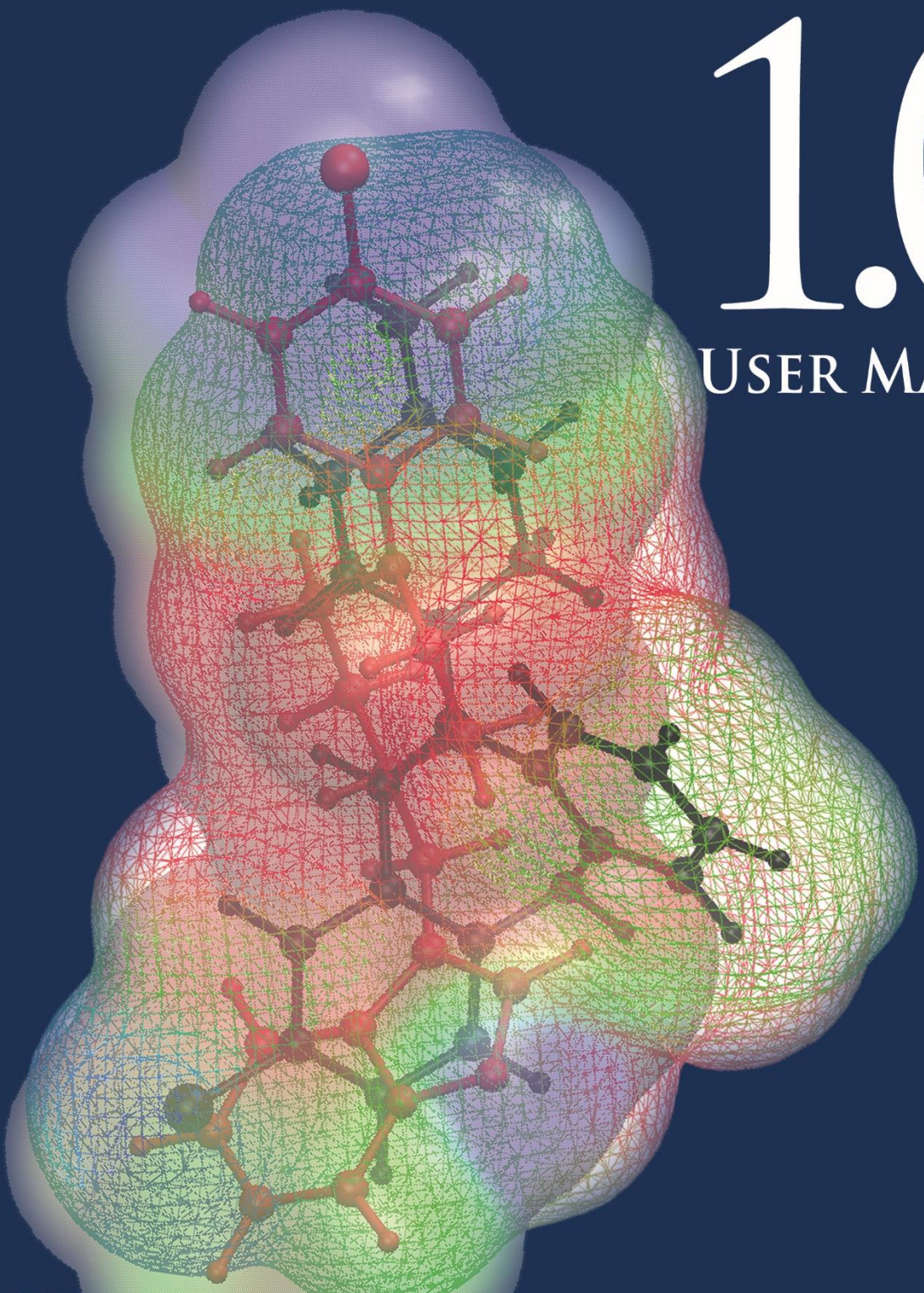
CEPOS INSILICO
SCIENCE FOR SCIENTISTS



1.0

USER MANUAL

CIMATCH™

Impressum

Waldstraße 15
90587 Obermichelbach
www.ceposinsilico.com

Manual Tim Clark

Layout **www.eh-bitartist.de**

CEPOS
CEPOS INSILICO GMBH

TABLE OF CONTENTS

ABOUT CIMATCH™

CImatch™ overlays molecules based on their standard isodensity surfaces and local properties projected onto them.

The simplest way to match the surfaces is a SHAPE match, which only considers the geometries of the two surfaces. **Figure 1** shows the algorithm schematically:

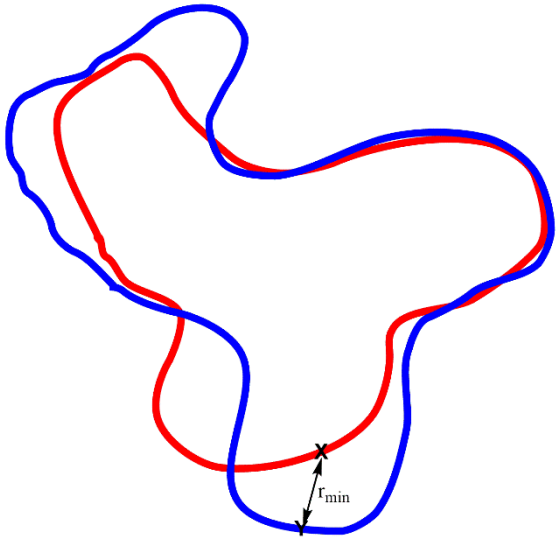


Figure 1 Schematic diagram of the algorithm used to match two surfaces

Similarity indices

The shape similarity index between surfaces A and B, S_{AB}^{R}, which is zero for identical surfaces (i.e. A = B), is given as:

$$S_{AB}^{R} = \sum_{i=1}^{N_A} r_{\min}^{i} \quad (1)$$

where N_A is the number of surface points for molecule A and $r_{\min}^{i}$ is the minimum distance from point i on surface A to any point on surface B. Note that any point on surface B (e.g. **Y** in **Figure 1**) may be closest to more than one point on surface A, and therefore occur several times for $r_{\min}^{i}$. This means that $S_{AB} \neq S_{BA}$.

The similarity index can be weighted by any of the local properties MEP (V), EA_L (A), IE_L (I) or α_L (P) to give the following similarity indices:

Shape	$S_{AB}^{R} = \sum_{i=1}^{N_A} r_{\min}^{i}$
Shape × MEP	$S_{AB}^{RV} = \sum_{i=1}^{N_A} r_{\min}^{i} \cdot \delta V_{\min}$
MEP	$S_{AB}^{V} = \sum_{i=1}^{N_A} \delta V_{\min}$
Shape × IE_L	$S_{AB}^{RI} = \sum_{i=1}^{N_A} r_{\min}^{i} \cdot \delta I_{\min}$
IE_L	$S_{AB}^{I} = \sum_{i=1}^{N_A} \delta I_{\min}$
Shape × EA_L	$S_{AB}^{RA} = \sum_{i=1}^{N_A} r_{\min}^{i} \cdot \delta A_{\min}$
EA_L	$S_{AB}^{A} = \sum_{i=1}^{N_A} \delta A_{\min}$
Shape × α_L	$S_{AB}^{R\alpha} = \sum_{i=1}^{N_A} r_{\min}^{i} \cdot \delta\alpha_{\min}$
α_L	$S_{AB}^{\alpha} = \sum_{i=1}^{N_A} \delta\alpha_{\min}$

In these cases, $\delta V_{\min}$ is the difference in the MEP projected onto the two surface points (**X** and **Y** in **Figure 1**), and analogously for the other surface properties. Note that for the similarity indices that do not include the shape, the points are still matched as shown in **Figure 1** but the distance between them does not enter the equation for the similarity index.

The closer the surface points and the smaller the difference between their local-property values, the better the match. A perfect match gives an overlay value of zero. Overlaying a molecule on itself is a good way to test program options to optimize performance. The program's default options are conservative and chosen to give reproducible results at the expense of calculational effort.

The matching algorithm

The program uses a genetic algorithm to find the best global overlay. The requested similarity index is optimized by translating and rotating molecule B while holding A constant. The optimization process consists of *repeat* iterations of complete evolution runs. Each run uses *npop* members of the population and runs for *maxgen* generations. The overall result is the lowest found in the series of evolution runs.

The size of the population (default 32), maximum number of generations (default 200) and number of repeat runs (default 20) are adjustable parameters.

CALLING CIMATCH™

```
<path> CImatch.exe <mol1> <mol2> <options>
```

CImatch™ is designed for efficient parallel execution; it will use all available cores.

Input files

The default input file type is the ParaSurf™ output file *<mol>.psf*. However, in the absence of the *force* option, CImatch™ will check which input files are available and, if necessary, perform ne EMPIRE™ and/or ParaSurf™ calculations. In this case, a file named *CImatch.com* in the current directory is needed to define the calls to EMPIRE™ and ParaSurf™ (see below).

NOTE: Giving the *.psf* extension is optional in the program call (i.e. *CImatch.exe mol1 mol2* and *CImatch.exe mol1.psf mol2.psf* are identical.

NOTE: *CImatch.exe mol1 mol2* is not the same as *CImatch.exe mol2 mol1*, as outlined above.

Options (case insensitive)

function=***<fn>***		*<fn>* defines the similarity index to be optimized. It may be one of:	
	R	Shape only (S_{AB}^{R})	
	RV	Shape × MEP (S_{AB}^{RV})	Default
	V	MEP (S_{AB}^{V})	
	RI	Shape × IE_L (S_{AB}^{RI})	
	I	IE_L (S_{AB}^{I})	
	RA	Shape × EA_L (S_{AB}^{RA})	
	A	EA_L (S_{AB}^{A})	
	RP	Shape × α_L ($S_{AB}^{R\alpha}$)	
	P	α_L (S_{AB}^{α})	
npop=***<n>***		Defines the size of the population in the genetic algorithm.	Default=32

maxgen=*<n>*		Defines the number of generations within a single optimization run.	Default=200
repeat=*<n>*		Defines the number of times the optimization should be repeated.	Default=20
force=*<s>*		Forces CImatch™ to request new EMPIRE™ and/or ParaSurf™ calculations	
	<s> = ***EMPIRE***	Do both new EMPIRE™ and ParaSurf™ calculations for template and target, regardless of which files are present.	
	<s> = ***PARASURF***	Only do new ParaSurf™ calculations for template and target, regardless of which files are present.	
mode=*<s1>*		Defines the precision with which the program works. Possible values are:	
	<s> = ***QAD***	“Quick and dirty” mode (npop=8, maxgen=100, repeat=3). This mode is intended for fast initial scans.	
	<s> = ***MED***	“Medium” mode (npop=24, maxgen=100, repeat=10). This mode is intended as a compromise between extensive searching and computational speed.	
	<s> = ***FIN***	“Fine” mode (npop=32, maxgen=200, repeat=20). This is the default mode and represents a very conservative setup for extensive sampling at the expense of cpu time.	Default
fragment=*<fragment_name>*		Requests that the fragment indicated as *<fragment_name>* in the input .sdf file be used as the template. (See the **ParaSurf™19 user manual** for details)	
mode=*<s1>,<s2>*		Requests that a first scan be made with multiple templates at the level *<s1>*, followed by selection of the best template conformations for calculations at level *<s2>*. The default option is given by *select* alone.	

select=*<i>*	Defines the number of templates to be selected.	Default is *<i>* = the larger of 3 or the number of targets/10

PROGRAM INPUT FILES

The minimum requirement for CImatch™ input (provided EMPIRE™ and ParaSurf™ are available) are *<mol1>.sdf* and *<mol2>.sdf* files for template and target as input for EMPIRE™. A second alternative starts with *<mol1>_e.sdf* and *<mol2>_e.sdf* output files from EMPIRE™ and only perfoms ParaSurf™ calculations. If *<mol1>.psf* and *<mol2>.psf* ParaSurf™ files are present, CImatch™ also requires that the two ParaSurf™ output SDF files *<mol1>_p.sdf* and *<mol1>_p.sdf* be present in the same directory as the input ParaSurf™ files. These SDF-files are required to write the CImatch™ output SDF file, which contains the overlaid geometries of the two molecules and their bonds, which are taken from the input ParaSurf™ SDF files. Both EMPIRE™ and ParaSurf™ SDF files may contain multiple molecules, which will all be processed.

The complete process to perform a CImatch™ overlay is therefore:

1. Perform EMPIRE™ single-point calculations or geometry optimizations on the two molecules to obtain EMPIRE™ *<mol>_e.sdf* files, which are used as input for

2. ParaSurf™, which calculates the input *<mol> .psf* files and updates the input *<mol> _e.sdf* files to *<mol>_p.sdf*. The default ParaSurf™ grid size for the surface is too small for efficient CImatch™ calculations, so that the option *mesh=0.5* should be used.

3. Perform the CImatch™ calculation as outlined above.

This process is performed automatically if the necessary files are not present. In this case, an addition file named *CImatch.com* in the working directory is necessary to define the calls to EMPIRE™ and ParaSurf™ (see below).

The CImatch.com command file

If EMPIRE™ and/or ParaSurf™ calculations are to be performed, the call for the two programs must be given in a file named *CImatch.com* in the working directory. For Windows, a typical *CImatch.com* might be:

```
c:\bin\empire_AM1spt.bat
c:\bin\parasurf_AM1.bat
```

where *empire_AM1spt.bat* is a script to run an AM1 single-point calculation with EMPIRE™ and *parasurf_AM1.bat* to run a subsequent ParaSurf™ calculation. It is important in the latter that the **ParaSurf™ options** *psf=on* and *mesh=0.5* are given.

PROGRAM OUTPUT

CImatch™ writes a text file *<mol1>_<mol2>_<FUNCTION>.log*, an SDF file of the overlaid molecules *<mol1>_<mol2>_<FUNCTION>.sdf* and one of the target in its overlaid coordinates *<mol2>@<mol1>_<FUNCTION>.sdf*.

The CImatch™ log file

The program log file provides details of program execution and the calculated similarity indices. **Box 1** shows a log file obtained using default parameters.

```
      CCC    III                   t           h    (TM)
     C   C    I   m m mm    aaa   ttt     cc   h
     C        I   mm m  m   a   a   t     c  c  h hh
     C        I   m  m  m       a   t     c     hh  h
     C        I   m  m  m   aaaa   t     c     h   h
     C        I   m  m  m  a   a   t     c     h   h
     C   C    I   m  m  m  a   a   t  t  c  c  h   h
      CCC    III  m  m  m  aaa a    tt    cc   h   h      (c) Cepos InSilico GmbH, Obermichelbach 2017, 2019

Started at: Fri Apr 26 15:56:51 2019 on PC_TIM


<> Template molecule         : 8_e
<> Target molecule           : 9_e
<> Target function           : Shape * MEP
<> Population                :     32
<> Maximum Nr. generations   :    200
<> Nr. evolutionary cycles   :     20
<> Numbers of surface points : Template:      672 Target:      704
<> Maximum shift             :    2.617 Angstrom
<> Translational resolution  :    0.021 Angstrom

Cycle  1:        1.997 : Best yet :       1.997
Cycle  2:        2.035 : Best yet :       1.997
Cycle  3:        2.019 : Best yet :       1.997
Cycle  4:        2.040 : Best yet :       1.997
Cycle  5:        1.932 : Best yet :       1.932
Cycle  6:        2.090 : Best yet :       1.932
Cycle  7:        2.032 : Best yet :       1.932
Cycle  8:        2.127 : Best yet :       1.932
Cycle  9:        2.043 : Best yet :       1.932
Cycle 10:        2.016 : Best yet :       1.932
Cycle 11:        1.956 : Best yet :       1.932
Cycle 12:        2.042 : Best yet :       1.932
Cycle 13:        2.012 : Best yet :       1.932
Cycle 14:        2.051 : Best yet :       1.932
Cycle 15:        1.999 : Best yet :       1.932
Cycle 16:        2.039 : Best yet :       1.932
Cycle 17:        1.954 : Best yet :       1.932
Cycle 18:        1.946 : Best yet :       1.932
Cycle 19:        1.982 : Best yet :       1.932
Cycle 20:        2.132 : Best yet :       1.932

<> Scores (Function optimized indicated by asterisks):

                                                          *********
      Shape        MEP        IE(l)       EA(L)       Pol(L)     Shape*MEP  Shape*IE(L) Shape*EA(L) Shape*Pol(L)
      0.355      3.673      19.685      22.928      99.602       1.932       8.169       8.096      35.012
```

Box 1 Example of a CImatch™ log file using default parameters and a single template and target

After printing the header and the details of the calculation, CImatch™ calculates the maximum translation allowed for the second molecular center relative to the first and gives the resolution of the translational moves. These two parameters are determined from the maximum dimensions of the molecules involved. The translational range is designed to cover all possible significant overlays and the resolution results from adapting this range to the overlay genes.

The program then prints the results of each evolutionary run (Cycle); in this case 20. The best result in all runs is stored and becomes the final output. Note that the results of the individual runs vary between 1.932 and 3.132, a typical range for multiple overlays of the same molecules

The final scores (i.e. those for the best solution) are given as a one-row table. The target parameter (in this case *Shape*MEP*) is marked by the row of nine asterisks. The scores for the other variables at the geometry of the optimized overlay for the target *Shape*MEP* are given in the other columns.

The CImatch™ target .sdf file

The orientation and position of the overlaid target molecule are written in a file named *<mol1>@<mol2>_<function>.sdf* with bonds taken from *<mol2>_e.sdf*.

The CImatch™ overlay .sdf file

CImatch™ writes an output .sdf file with the two molecules in the calculated overlay. The bonds of the two individual molecules are taken from the input *<mol>*_e.sdf files.

Figure 2 shows a visualization of such an overlay:

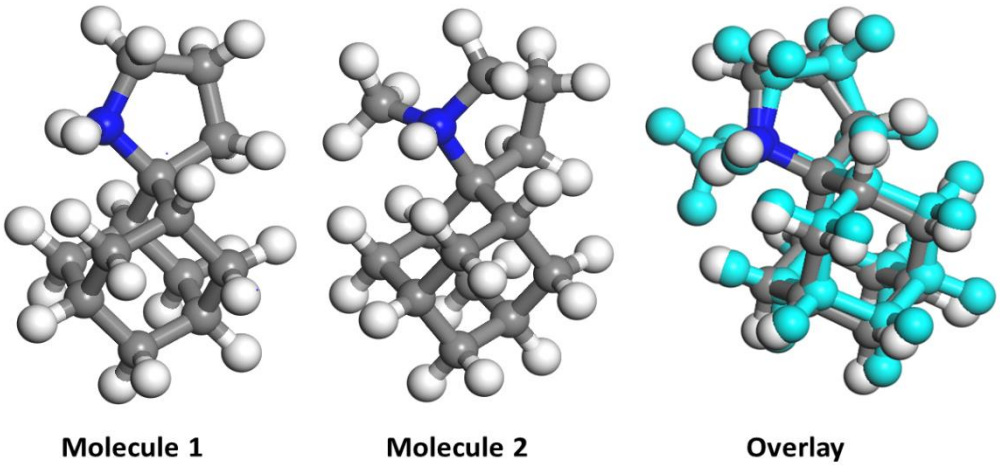


Figure 2 The two input molecules and the resulting overlay from the CImatch™ calculation shown in **Box 1**

The CImatch™ table .csv file (*Template_<input1>_<function>.csv*)

CImatch™ writes an output *.csv* file that lists the results for all templates and targets the separator is a comma). Separate tables of all the targets are written for multiple templates and for two-stage runs an additional table for the selected targets at the second calculation mode. This file can be imported directly into spreadsheet programs.

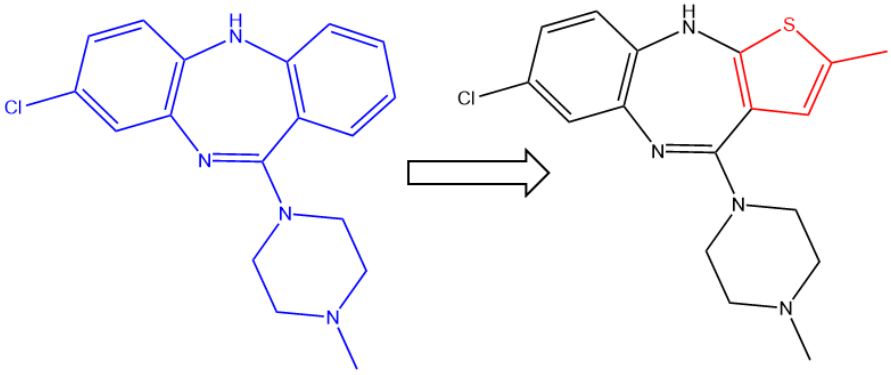
Using CImatch™ for substructure searches, finding bioisosteres and scaffold hopping

If a molecular fragment is used as the template in CImatch™, the closest equivalent to the fragment will be found in the target molecule. All that is necessary is to add a "*> <FRAGMENTS>*" section to the input .sdf file for the template and to define the fragment in the CImatch™ call. Details of defining fragments for ParaSurf™ are given in the **ParaSurf19™ manual**. For fragment "*fragment_name*" in the SDF-input file "*mol1.sdf*", the CImatch call is:

```
CImatch.exe mol1 mol2 fragment=fragment_name
```

Note that fragments are only recognized for the template molecule, not for targets.

Figure 3 shows the results of a fragment overlay for example given in **Wikipedia for bioisosteres**.



The red methyl-thiophene fragment, which is a non-classical bioisostere for the benzo group, was overlaid on the blue molecule using the RV function. The search was not constrained but the overlay correctly identifies the benzo group, rather than, for instance the chloro-benzyl, as the bioisostere. The calculated S_{AB}^{RV} is 1.3 Å kcal mol^{-1}, indicating high similarity.

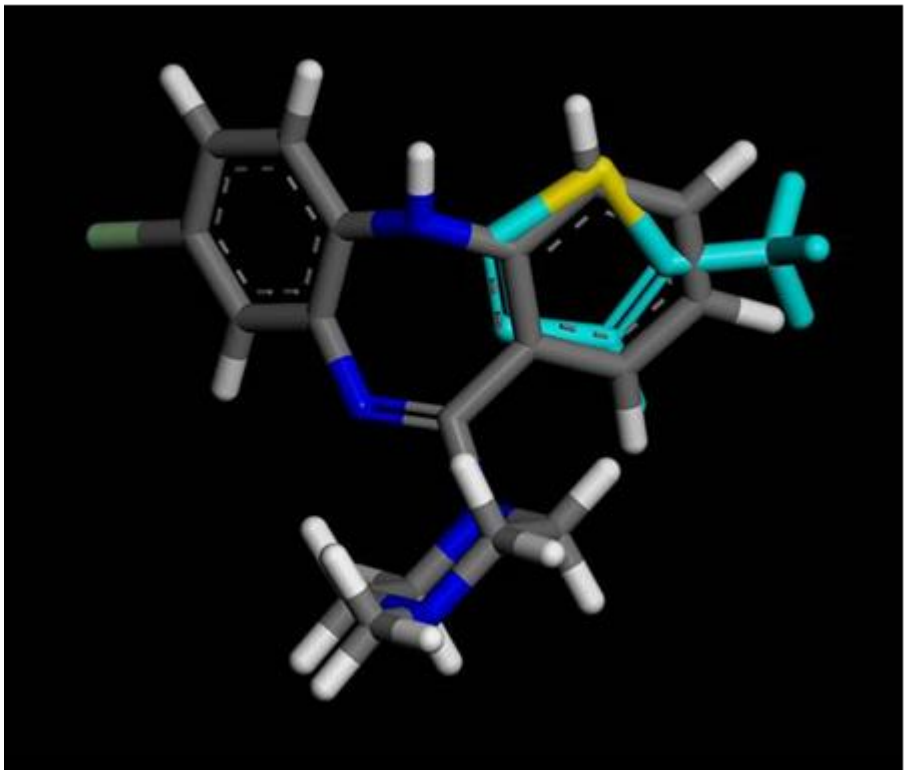


Figure 3 Overlay of the red methyl-thiophene fragment on the blue molecule shown above. The calculation used the AM1 Hamiltonian and the *shape × MEP* target function. The overlay clearly identifies the bioisostere.

SUPPORT

Contact

Questions regarding CImatch™ should be sent directly to:

support@ceposinsilico.com

Cepos InSilico GmbH

Waldstraße 15
90587 Obermichelbach
Germany

support@ceposinsilico.com
Tel. +49 (0)9131 970 4910
Fax. +49 (0)9131 970 4911

www.ceposinsilico.com/contact

LIST OF FIGURES AND BOXES